(12)

12

HIERARCHICAL BAYES MODELS FOR THE PROGRESSION OF HIV INFECTION USING LONGITUDINAL CD4$^+$ COUNTS

by

Nicholas Lange

Bradley P. Carlin

Alan E. Gelfand

TECHNICAL REPORT No. 461

NOVEMBER 27, 1992

Prepared Under Contract

N00014-92-J-1264 ((NR-042-267)

FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA 94305-4065

HIERARCHICAL BAYES MODELS FOR THE PROGRESSION OF HIV INFECTION USING LONGITUDINAL CD4$^+$ COUNTS

by

Nicholas Lange

Bradley P. Carlin

Alan E. Gelfand

TECHNICAL REPORT No. 461

NOVEMBER 27, 1992

Prepared Under Contract

N00014-92-J-1264 ((NR-042-267)

FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA 94305-4065

Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4$^+$ counts

Nicholas Lange Bradley P. Carlin Alan E. Gelfand*

Taking the absolute number of CD4$^+$ cells (also known as T helper cells, T4 cells, and CD4 cells) as a marker of disease progression for persons infected with the human immunodeficiency virus (HIV) we model longitudinal series of such counts for a sample of 331 subjects in the San Francisco Men's Health Study. We conduct a careful and fully Bayesian analysis of these data. We are able to employ individual level nonlinear models incorporating critical features such as incomplete and unbalanced data, population covariates, unobserved random change points, heterogeneous variances, and errors-in-variables. Using results of previously published work from several different sources we construct rather precise prior distributions. Our analysis provides marginal posterior distributions for all population parameters in our model for this cohort. Using an inverse prediction approach we also develop the posterior distributions of time for CD4$^+$ count to reach a specified level.

KEY WORDS: AIDS, Gibbs sampler, growth curves, heterogeneity, inverse prediction, marginal posterior distribution, prior specification, random change points, sexual preference.

1. Introduction

A tremendous amount of published work has appeared in the last few years on the measurement of various clinical markers of disease and disease progression in persons infected with the human immunodeficiency virus (HIV). Among the primary biological markers of interest is the number of $CD4^+$ cells per cubic millimeter of drawn blood (Fauci et al. 1984; Bowen et al. 1985; Eyster et al. 1985, among many others). These cells have also been called CD4 cells, T4 cells, and T helper cells in the immunologic literature. This paper deals solely with this clinical marker. Although the need for a full and detailed understanding of the behavior of $CD4^+$ counts over time is obvious (and is motivated in detail in the next section) little of this published work has set out to model the longitudinal processes involved in an adequate manner. We feel that this lack is due primarily to the urgent initial clinical need to establish useful critical values of this marker (for example $\leq$ 500 cells/mm^3, or $\leq$ 200 cells/mm^3) but also to the statistical difficulties of modeling the process even when a longitudinal data set (i.e. repeated $CD4^+$ counts for the same individual over time) has been obtained.

In a recent paper, De Gruttola, Lange and Dafni (1990) addressed some of the statistical modeling issues and offered results that help meet clinical needs. Our investigation goes further in meeting such needs through the removal of somewhat restrictive and implausible modeling assumptions required to implement classical frequentist and empirical Bayes techniques, while making more use of available data and prior evidence. We are able to accommodate individual level nonlinear models incorporating critical features such as incomplete and unbalanced data, population covariates, unobserved random change points, heterogeneous variances, and errors-in-variables. We construct rather precise prior distributions due to the availability of previously published work. We are able to obtain marginal posterior distributions of interest using recent advances in the technology of Bayesian computations. In total we provide clinically useful results based on a full and careful statistical analysis of a current data set along with information gleaned from previously published analyses of relevant data.

The structure of this paper is as follows. The next section gives a brief review of the medical, epidemiological and biostatistical literature on $CD4^+$ counts and on our data set in particular. Section 3 gives detail on our high-dimensional model and on our prior specifications drawn from the sources outlined in section 2. In section 3 we also define and propose a solution to an inverse prediction problem in our context. The final sections 4 and 5 give results, conclusions and discussion.

2. Background

The following is a brief but fairly complete synopsis of relevant published literature in the analysis and clinical uses of $CD4^+$ counts for HIV infected individuals, a motivation of the need for more data-analytic results in this area and a description of the major new and necessary features of our modeling approach.

Eyster et al. (1985) found a decrease in the absolute number of $CD4^+$ cells over time in HIV-infected hemophiliacs for whom a time of seroconversion (development of detectable antibodies to HIV) could be determined. In further studies, Eyster et al. (1987) and Eyster et al. (1989) found that a low $CD4^+$ count had a high sensitivity and high predictive value for the development of the acquired immunodeficiency syndrome (AIDS). These authors also defined a normal range of $CD4^+$ cell counts for uninfected individuals of $500 - 1100$ cells/mm^3 (cf. their Figure 2). A study by Devash et al. (1990) used $464 - 1364$ cells/mm^3 as normal levels (cf. their Table 2). Volberding et al. (1990) used a value of $CD4^+$ count ≤ 500 at study entry as an eligibility criterion for patient selection in a randomized clinical trial for zidovudine (formerly AZT) involving 1338 subjects. One measure of the clinical effectiveness of zidovudine in this study was an observed increase in the numbers of $CD4^+$ cells in the treated group relative to the control group, in which they observed an average decrease of 16 $CD4^+$ cells/year. Phair et al. (1990) found that $CD4^+$ counts ≤ 200 greatly increased the risk of an AIDS-related disease (*Pneumocystis carinii* pneumonia, or PCP) involving 1665 HIV-infected subjects. For additional results on relationships between PCP risk and prophylaxis, HIV infection and $CD4^+$ cell count, see also the Centers for Disease Control (1989), Masur et al. (1989), Leibovitz et al. (1990) and Leoung et al. (1990). For similar uses of $CD4^+$ counts as eligibility criteria and secondary measures, see Fischl et al. (1990) and Collier et al. (1990).

Taylor et al. (1989) provided an analysis of the statistical effectiveness of various functions of lymphocyte counts. These authors noted a high degree of within-person variability in these counts regardless of their functional representation. They also noted that the ratio of $CD4^+$ cells to the total number of lymphocytes, when available, may have slightly better prognostic significance than the absolute number, due to its slightly lower variability within individuals. In a further study of eight different biological markers, Fahey et al. (1990) concluded that progression to AIDS was predicted most accurately by the level of $CD4^+$ cells (in absolute number or as a ratio) in combination with two serologic measurements.

Although $CD4^+$ cell count has been well established as an important clinical marker in studies of the progression of HIV infection, further understanding of its behavior from seroconversion to the onset of AIDS is required for its use in analyses of the efficacies of medical interventions, in patient counseling, palliative care and AIDS hospice studies, and in the design and analysis of clinical trials for anti-HIV therapies. For instance, in counseling an asymptomatic HIV-infected individual who, on the basis of a single $CD4^+$ measurement between 200 and 500, is found to be in a high-risk category for further progression towards clinical AIDS or of contracting PCP, it would be extremely useful to know how long it may take for the $CD4^+$ cell count for this individual to drop below 200, or to below 100, for then the onset of AIDS may be much closer. (Obtaining expected time of death from AIDS, however, is a very different matter and is not dealt with in this paper.)

For many of the motivating purposes given at the beginning of the preceding paragraph there remain clear needs for proposing and analyzing useful and adequate statistical models for the stochastic processes generating observed $CD4^+$ cell counts from the time of infection with HIV onwards. To these ends, and due to the availability of much prior information on the behavior of $CD4^+$ cell counts during this time interval, we have chosen to use a family of parametric growth-curve models in a fully Bayesian analysis of repeated $CD4^+$ counts for asymptomatic HIV-infected individuals. We describe our particular models and prior constructions in the next section. Bayesian analyses of linear growth-curves have been given previously by Geisser (1970), Lee and Geisser (1972), Fearn (1975) and Rao (1987, Section 2, as well as references therein), among others (see also Cox and Solomon, 1986, for several alternate parametric frequentist approaches). Anticipating problems of regression to the mean in repeated-measures studies, these models include individual level random effects that allow for random deviations of an individual's trajectory from the expected trajectory for the population to which the individual belongs. By allowing for random effects, such as individual random intercepts and individual random slopes, these models simplify the specification of the dependence structure among the serial observations when compared with more classical multivariate approaches. This feature of the parametric growth-curve models is particularly attractive in the present context due to the high variability of $CD4^+$ counts both within and between individuals and the relatively small numbers of repeated measurements available for each subject compared to the large number of subjects available.

In addition to problems of high variability and limitations due to dealing only with large numbers of short time series, certain estimators of mean $CD4^+$ counts and of changes in these counts over time can be seriously biased. Subjects are often available for study only long after they have become

infected. Brookmeyer and Gail (1987) have pointed out the need to accommodate the differing and typically unknown times between infection and study entry when estimating the effects of covariates on the risk of disease using prevalent cohorts for new diseases such as AIDS. These authors label this potential for bias "onset confounding," and state (p. 745) that no reliable inferences can be drawn from prevalent cohorts if onset confounding is present. This disappointing conclusion is certainly valid if no external source of data is available for estimating the distribution of the potential biasing factor. De Gruttola et al. (1990) have conducted a complete-case analysis of repeated $CD4^+$ counts obtained by the San Francisco Men's Health Study (SFMHS) for 201 HIV-infected men sampled from a high-risk group, each subject observed at five six-month intervals beginning in 1984. These authors addressed the problem of bias from onset confounding by using an external estimate of the infection-time distribution calculated by Bacchetti and Moss (1989).

This paper presents a careful Bayesian analysis of the SFMHS data using prior distributions developed from previously reported similar studies. We contribute to the understanding of $CD4^+$ cell count progression by generalizing the models and results given by De Gruttola et al. (1990) in several important and essential ways, as described in the next section. The SFMHS data we have available for our study consist of up to five repeated measurements of $CD4^+$ counts in blood drawn at six-month intervals beginning in 1984 for 331 male subjects sampled from contiguous neighborhoods in San Francisco. All subjects have been infected with HIV for unknown periods of time. As is customary, we take infection time to be the time of seroconversion. The small group of 27 individuals who seroconverted after study entry were excluded from the analysis. For each subject we also have available his age at study entry (in years) and his self-reported sexual preference (homosexual or bisexual). The small number of men in neither of these two sexual preference categories were also excluded from the analysis.

3. Model specifications

In this section we first note the novel features of our longitudinal model for $CD4^+$ cell counts. Since our approach is fully hierarchical Bayesian, we provide detail on both its likelihood and prior specifications. In particular we elucidate the nature of the information and considerations that give rise to the latter. We employ the Gibbs sampler for hierarchical Bayes modeling (see Gelfand and Smith, 1990; Gelfand et al., 1990, for details) to obtain desired marginal posterior distributions. The Gibbs sampling approach is attractive in that it is a computational technique that can readily handle our very high-dimensional model

with more than 2,000 parameters. Complete conditional distributions required for the Gibbs sampling technology are also presented. Finally, we discuss the problem of inverse prediction, i.e. within the Bayesian framework how does one predict time to reach a pre-specified $CD4^+$ cell count?

3.1. Model features

We propose a longitudinal model for the $CD4^+$ cell counts for the subjects in the San Francisco Men's Health Study that incorporates the following features:

(i) a growth curve for the counts that is nonlinear over time. In fact we employ a piecewise-linear curve of two pieces motivated by the work of Lang et al. (1989) and Eyster et al. (1989); see also Masur, et al. (1989). These studies suggest a rather slow decline in $CD4^+$ counts in the early years post-seroconversion, followed by a more precipitous decline during the few years prior to diagnosis of AIDS;

(ii) repeated measurements on individuals resulting in incomplete structure due to varying numbers of measurements (at least one and at most five) per individual. Dependence between individual measurements is captured through individual level random effects;

(iii) error structure that is assumed normal but with nonhomogeneous variances. That is, given the random effects we assume that the individual level errors are independent Gaussian but avoid the implausible assumption of a common variance shared by all subjects;

(iv) inclusion of covariate effects for sexual preference (self-reported homosexual, bisexual) and age (in years). Brookmeyer and Goedert (1989) demonstrated dependence between a binary indicator for age (either < 20 or ≥ 20 years) and $CD4^+$ count in their study of a hemophiliac cohort. We study this dependence further by treating age as a continuous covariate. Were additional covariates available, particularly ones reflecting individual comorbidities, we would have included them as well;

(v) an errors-in-variables component to adjust for unknown random infection times prior to study entry and data collection in 1984.

Our model, which we detail in the following subsection, can be given in a general form which, by virtue of accommodating nonlinearity, incomplete and unbalanced measurements, nonhomogeneous errors and errors-in-variables, extends the setting provided by Lange and Laird (1989). Fully Bayesian analysis for this class of models is straightforward though computationally demanding, following the development of sections 3.2–3.5.

3.2. Model detail and notation

Let Y_{ij} be the (possibly unobserved) CD4$^+$ cell count for the ith subject at the jth sampling period, $i = 1, \ldots, n; j = 0, 1, 2, 3, 4$. The five individual samples were taken semi-annually thus spanning a total of two years. These times are denoted by Z_j measured in years, i.e., $Z_0 = 0, \ldots, Z_4 = 2.0$. For many subjects in the data set a CD4$^+$ cell count is missing for at least one time period.

Since we model only those subjects who seroconverted prior to Z_0, each individual is assumed to have a random unknown offset τ_i which is the time from seroconversion to the start of the study. Hence for the ith subject the actual time from seroconversion to the jth sampling period is $Z^*_{ij} = Z_j + \tau_i$. Modeling an explanatory variable in this fashion is typically referred to as an errors-in-variables approach. However, here the τ_i are not assumed to be random errors about zero; in fact, $\tau_i > 0$ with probability one.

At the individual level, suppose that a piecewise linear growth curve with two pieces describes expected CD4$^+$ cell counts over time. Figure 1 provides an illustration. The first line segment, $\eta_{0i}+\eta_{1i}Z$, shows slow decline in expected CD4$^+$ cell count for smaller Z. The second line segment, $\eta_{2i} + \eta_{3i}Z$, shows more rapid decline for larger Z, after some unknown time point t_i, presaging a diagnosis of AIDS within a few years. Lang et al. (1989, Figure 1, page 66) suggest that such overall behavior of CD4$^+$ cell counts can be expected to begin at roughly six months after seroconversion.

Since the line segments agree at t_i, there are, in fact, only three distinct parameters. We simplify notation by letting $\Delta_i = \eta_{1i} - \eta_{3i}$, hence $\Delta_i t_i = \eta_{2i} - \eta_{0i}$, and thus write

$$Y_{ij} = \eta_{0i} + \eta_{1i} Z^*_{ij} - \Delta_i \left(Z^*_{ij} - t_i\right)^+ + \epsilon_{ij}, \tag{1}$$

where $c^+ = \max(c, 0)$. In (1) the $\epsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma_i^2\right)$ independent of the random η_{0i}, η_{1i} and Δ_i, which are assumed independent of each other as well. We model η_{0i}, η_{1i} and Δ_i as follows:

$$\begin{aligned} \eta_{0i} &= \alpha_0^o + \alpha_0^{\mathrm{A}} a_i + \alpha_0^{\mathrm{S}} b_i + \beta_{0i}\,; \\ \eta_{1i} &= \alpha_1^o + \alpha_1^{\mathrm{A}} a_i + \alpha_1^{\mathrm{S}} b_i + \beta_{1i}\,; \\ \Delta_i &= \alpha_2^o + \alpha_2^{\mathrm{A}} a_i + \alpha_2^{\mathrm{S}} b_i + \beta_{2i}\,. \end{aligned} \tag{2}$$

The nine αs in (2) denote population effects, while the three βs denote individual level effects. Thus with τ_i, t_i and σ_i^2 there are six individual parameters for each subject. For our sample of $n = 331$ subjects we have a total of $6(331) + 9 = 1995$ parameters, compared with 1436 data points.

The reader may be troubled by a model involving 1995 parameters with only 1436 data points and perhaps be inclined to conclude that the model will "swamp" the data and drive conclusions. In

fact, there are only 9 population level parameters with the remaining individual level parameters of lesser interest. Moreover, the exchangeability assumption in our hierarchical Bayes model (see section 3.3) is its crucial aspect. Exchangeability enables one to employ information from the entire collection of subjects to strengthen the inference for any individual subject. (This "borrowing strength" idea is implicit in the construction of "shrinkage estimators" which pull the individual level estimates toward a population level estimate.) We understand any individual's growth curve better (at a micro-level) because of the longitudinal data gathered from other similar subjects. We can also then infer average or population level growth curve behavior effectively (at a macro-level) by appropriate aggregation.

Each vector $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})^{\mathrm{T}}$ is assumed, as is customary, to be normally distributed, i.e. $\beta_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{V})$, $\mathbf{V}$ unknown. The $\alpha_p^o, p = 0, 1, 2$, denote population baseline effects for intercept, slope and change in slope respectively. With a_i denoting the age (in years) of the ith subject, the $\alpha_p^A, p = 0, 1, 2$, scaled age effects that adjust respective baseline effects. Similarly, with b_i denoting sexual preference ($b_i = 0$ for self-reported homosexual, $b_i = 1$ for self-reported bisexual), the $\alpha_p^{\mathrm{S}}, p = 0, 1, 2$, provide sexual preference effects that also adjust respective baseline effects.

It is easier to work with a model involving only the observed counts; hence we alter our notation slightly. If the ith subject is observed on s_i occasions at times $Z_j^o, j = 1, \ldots, s_i, \{Z_1^o, \ldots, Z_{s_i}^o\} \subset \{0, 0.5, 1.0, 1.5, 2.0\}$, let $\mathbf{Y}_i^o$ denote the associated vector of observed CD4$^+$ cell counts. Furthermore let $\mathbf{Y}^o$ denote the concatenated vector of all observed counts for all subjects. In addition, let

$$\alpha^o = \begin{bmatrix} \alpha_0^o \\ \alpha_1^o \\ \alpha_2^o \end{bmatrix}, \alpha^{\mathrm{A}} = \begin{bmatrix} \alpha_0^{\mathrm{A}} \\ \alpha_1^{\mathrm{A}} \\ \alpha_2^{\mathrm{A}} \end{bmatrix}, \text{ and } \alpha^{\mathrm{S}} = \begin{bmatrix} \alpha_0^{\mathrm{S}} \\ \alpha_1^{\mathrm{S}} \\ \alpha_2^{\mathrm{S}} \end{bmatrix}.$$

We condense notation further by letting α be the 9×1 concatenated vector of these baseline and covariate-specific mean effects. Finally, let β denote the collection of all β_i, σ^2 the collection of all σ_i^2, τ the collection of all τ_i and $\mathbf{t}$ the collection of all t_i.

Thus for all n subjects the conditional likelihood (given β) becomes

$$\mathcal{L}(\alpha, \beta, \sigma^2, \tau, \mathbf{t}; \mathbf{Y}^o) = \prod_{i=1}^{n} \prod_{j=1}^{s_i} \mathcal{N}(Y_{ij}^o | \mu_{ij}, \sigma_i^2), \tag{3}$$

where μ_{ij} is the righthand side of (1) apart from ϵ_{ij}. (We use $\mathcal{N}(a_1|a_2, a_3)$ to indicate that the variable a_1 is a Gaussian random variable with mean a_2 and variance a_3, and employ similar notation for the other distributions specified.) The β_i are assumed i.i.d., so their joint distribution is

$$\prod_{i=1}^{n} \mathcal{N}(\beta_i | \mathbf{0}, \mathbf{V}). \tag{4}$$

3.3. *Prior specification*

We adopt a Bayesian analysis of the model given by (1)-(4) for several reasons. First, classical analysis for such an unbalanced and incomplete nonlinear random-effects model with nonhomogeneous variances is virtually impossible to carry out exactly, while analysis based on asymptotic theory seems unjustifiable in view of the small number of repeated measurements per subject. Second, just as a frequentist assumes the β_i exchangeable it seems plausible to think of the σ_i^2, τ_i and t_i exchangeable as well and to take advantage of this in the modeling process. Third, several previous studies provide substantial prior information with regard to the parameters α^o, τ_i and t_i that ought to be used. Fourth, consider the remaining parameters $\beta_i, \alpha^A, \alpha^S$ and σ_i^2. While we have no direct prior information for these unknowns, we have indirect information pertinent to their prior specification arising from natural centering along with suitable scaling based on the known ranges of the Y_{ij}^o, a_i and b_i. Again, use of such information enhances the modeling process. To address concerns about prior sensitivity we employ a more vague prior specification for these parameters; discussion of the data analysis in section 4 shows that our findings are robust to such changes. Fifth, recent advances in methodology for Bayesian computations enable reasonably straightforward implementation of a fully Bayesian analysis for such a challenging high-dimensional model.

Our prior information is drawn primarily from the work of Bacchetti and Moss (1989) and Lang et al. (1989). The former provides discussion useful for modeling the τ_i and the t_i while the latter is informative for modeling α and the t_i. Recalling the definition of τ_i (see Figure 1 and accompanying discussion) we seek information on the distribution of seroconversion times for the subjects in our study. Bacchetti and Moss (1989) provided an estimate of this distribution developed from a clinical cohort study at the San Francisco General Hospital. This distribution seems an appropriate prior for the subjects in our data set as they, too, are a group of San Francisco males. The Bacchetti and Moss estimate begins at January 1978. Since we consider only those subjects who have already seroconverted by the start of data collection in 1984, we truncate this distribution at six years. Rather than assuming an arbitrary parametric form for this distribution we instead divide this roughly six-year span into twelve intervals, each of length six months, and label each interval by its midpoint. Thus we assume the τ_i are i.i.d. following a twelve-point distribution taking on values $1/4, 3/4, \ldots, 23/4$. Adapting the Bacchetti and Moss estimate we obtain the prior distribution for the τ_i shown in Table 1. As an aside, the ensuing development shows that refinement of this prior to more than twelve points can be handled easily. In fact, initially we used a six-point prior which yielded findings very similar to those reported in the next

section under the twelve-point distribution given here.

Next, recall the definition of t_i (see Figure 1 and accompanying discussion). By definition, t_i is a time point within the AIDS latency (equivalently incubation) period, this period being defined as the time from seroconversion to diagnosis of AIDS. Several studies have investigated this latency distribution, taking parametric, semi-parametric and nonparametric approaches, including Medley et al. (1987), Lui et al. (1988), Kalbfleisch and Lawless (1988), Bacchetti and Moss (1989), Longini et al. (1989), Lagakos and De Gruttola (1989) and Lifson et al. (1989). All of these studies are based on at most eleven years of data and they vary widely in their predictions. They suggest median latency times of seven to ten years and that roughly 60 to 80% of latencies will occur by the end of the tenth year. According to Lang et al. (1989), t_i occurs approximately two to three years prior to an AIDS diagnosis. See also Figure 4 of Masur et al. (1989), which exhibits a similar pattern of progression before development of PCP. Thus we consider two choices of discrete prior distribution for t_i in years, corresponding to the 20% and 40% collapsed tails as shown in Table 2. These distributions have medians at five and six years respectively. As with the τ_i, the t_i will be assumed i.i.d.

For α, a nine-dimensional multivariate normal prior was selected. We require specification of its prior mean, which we denote by the 9×1 vector $\mathbf{c}$. Previous studies provide information for the mean of the prior baseline effects, but not for prior means on age and sexual preference effects, which we thus set to zero. We assume the prior variance-covariance matrix has the block-diagonal form

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}^o & 0 & 0 \\ 0 & \mathbf{D}^A & 0 \\ 0 & 0 & \mathbf{D}^S \end{bmatrix},$$

implying that covariate effects for age and sexual preference are independent of the baseline effects and of each other.

According to the study of Lang et al. (1989), the mean number of CD4+ cells was 716 at six months after seroconversion (the time after which we propose the model in Figure 1). We take this value as the mean for the baseline intercept α_0^o. This study also reports a decline in mean CD4+ count of 42 cells per six month period in the early HIV seropositive stages. Thus we take -84 as the mean for the first line segment (pre-t_i) slope α_1^o (since Z is measured in years). Similarly, a decline in mean CD4+ count from 430 to 190 is reported in the eighteen months prior to AIDS diagnosis. Thus the mean slope for the second (post-t_i) line segment is taken to be -160, yielding a mean for the change in slope α_2^o to be 76. Lang et al. (1989) also provided standard errors associated with

mean CD4$^+$ count at various post-seroconversion time points. The largest of these is 69, so to be conservative we chose 70 as the standard error for α_0^o. We approximate the variance of a slope by $\text{var}\left((Y_2^o - Y_1^o)/(t_2 - t_1)\right) \approx (t_2 - t_1)^{-2}(\text{var}(Y_2^o) + \text{var}(Y_1^o))$, where t_1 and t_2 are two time points on the same line segment and Y_1^o and Y_2^o are corresponding observed CD4$^+$ cell counts at these time points. For α_1^o, using the standard errors from Lang et al. (1989) of 15 and 17, three years apart, the approximation is $((15)^2 + (17)^2)/(3)^2$. Similarly, the variance of the post-t_i line segment slope is approximated using standard errors 28 and 25, eighteen months apart, yielding the approximation $((28)^2 + (25)^2)/(1.5)^2$. The variance for α_2^o is approximated by the sum of these two approximations.

It seems that the baseline α_p^o should be correlated. For instance, if α_0^o were decreased we would expect that α_1^o would increase to yield a line segment that continues to agree roughly with our prior information at particular time points. Similar arguments apply for α_1^o and α_2^o and for α_0^o and α_2^o. For simplicity we assume *a priori* that $\text{corr}(\alpha_0^o, \alpha_1^o) = -0.5, \text{corr}(\alpha_1^o, \alpha_2^o) = 0.5$, and $\text{corr}(\alpha_0^o, \alpha_2^o) = -0.5$. With these correlation approximations, the prior variance-covariance matrix $\mathbf{D}^o$ for α^o is thus determined.

Turning to the remaining parameters, we take prior specifications consistent with known ranges. Interpreting, for example, α^S as an adjustment to α^o and recalling the scales for α^o from the preceding paragraph, we suggest that plausible standard errors for α_0^S, α_1^S and α_2^S might be 100, 25 and 25 respectively. Assuming correlation structure as in $\mathbf{D}^o$ thus determines the variance-covariance matrix $\mathbf{D}^S$ for α^S. Similar reasoning leads to a variance-covariance matrix $\mathbf{D}^A$ for α^A, but since α^A enters (1) multiplied by age we rescale these standard errors by the average age of the subjects which is 36.1 (36.3 if homosexual, 35.1 if bisexual). Hence prior standard errors for α_0^S, α_1^S and α_2^S are $100/36.1, 25/36.1$ and $25/36.1$ respectively.

The σ_i^2 are assumed i.i.d. from an inverse gamma ($\mathcal{IG}$) distribution having mean μ = standard error $\sigma = (200)^2$. The inverse gamma distribution is parametrized customarily by λ_1 and λ_2, its shape and scale parameters respectively, from which one obtains $\mu = 1/(\lambda_2(\lambda_1 - 1))$ and $\sigma^2 = 1/(\lambda_2^2(\lambda_1 - 1)^2(\lambda_1 - 2))$. Finally, for $\mathbf{V}$ we choose an inverse Wishart distribution, i.e. $\mathbf{V}^{-1} \sim \mathcal{W}((\rho\Lambda)^{-1}, \rho)$. Recalling that $\mathbf{V}$ is the common variance-covariance matrix of the individual random effects, since Λ is roughly its prior mean we take Λ diagonal with diagonal elements $(25)^2, (6)^2$ and $(6)^2$ respectively. We make this prior rather vague by choosing a somewhat small precision of $\rho = 2$. We note in passing that if there was prior evidence to suggest that different population subgroups had different $\mathbf{V}$s, we could have easily included that possibility at this stage.

With the preceding priors, and assuming the $\beta_i, \sigma_i^2, \tau_i$ and t_i independent of each other and α, our overall model becomes

$$\mathcal{L}\left(\alpha, \beta, \sigma^2, \tau, \mathbf{t}; \mathbf{Y}^o\right) \cdot \mathcal{N}\left(\alpha | \mathbf{c}, \mathbf{D}\right) \cdot \prod_{i=1}^{n} \mathcal{N}\left(\beta_i | \mathbf{0}, \mathbf{V}\right) \cdot$$
$$\prod_{i=1}^{n} \mathcal{IG}\left(\sigma_i^2 | \lambda_1, \lambda_2\right) \cdot \prod_{i=1}^{n} p(\tau_i) \cdot \prod_{i=1}^{n} p(t_i) \cdot \mathcal{W}\left(\mathbf{V}^{-1} | (\rho\Lambda)^{-1}, \rho\right). \tag{5}$$

Numerical values in (5) for $\mathbf{c}, \mathbf{D}, \rho$ and Λ, and the distributions $p(\tau_i), p(t_i)$ and $\mathcal{IG}(\sigma_i^2 | \lambda_1, \lambda_2)$ have been specified in the preceding discussion. The dependence structure associated with (5) is captured succinctly in the directed graph shown in Figure 2. This graph provides a visual description of the structure of our problem. The "$\longrightarrow$"s indicate that all parameters (save $\mathbf{V}$) are connected to one another directly through the observed data. At each node (parameter) of this graph sits a "complete conditional distribution" (described in the next section). The "Gibbs sampler" idea (see Geman and Geman, 1984) is to sample from Markovian updates of these distributions in an iterative fashion resulting in random states for the graph whose joint distribution converges to the exact joint distribution of its nodes.

3.4. Complete conditional distributions

The Gibbs sampler offers a straightforward approach for obtaining marginal posterior distributions for the parameters in hierarchical Bayes models (see Gelfand and Smith, 1990; Gelfand et al., 1990). In the present case, the parameters of primary interest are the population parameters α. We may have some interest in summarizing individual level parameters as well.

The Gibbs sampler is an iterative Markovian updating scheme. We do not review details here, except that its implementation requires sampling from so-called complete conditional distributions, as noted in the previous subsection. For the remainder of this subsection we develop these distributions. Sampling is conducted with m parallel and independent replications each taken to r iterations. Desired marginal density estimates and features of these densities are obtained as Monte Carlo integrations of the corresponding complete conditional densities and features using the m replicates. Choice of m determines how close our marginal posterior density estimate is to the exact density at the rth iteration, while choice of r determines how close the latter density is to the actual marginal posterior density. Settings for m and r to achieve smooth converged estimates vary with the application. In the present situation we used $m = 500$ and $r = 25$. We remind the reader that each iteration for each replication required the generation of 2001 variates. Thus in total more than 2.5×10^7 variates were generated, with

more than 10^6 variates retained at the end of an iteration. Run time for all sampling and summaries on a single DEC station 3100 was roughly four hours wall clock time.

In order to make notation more compact at the individual level, let $\mathbf{w}_{ij} = [\,1 \quad Z^*_{ij} \quad -\left(Z^*_{ij} - t_i\right)^+\,]$, $\mathbf{W}_i = (\,\mathbf{w}_{i1}^{\mathrm{T}} \quad \ldots \quad \mathbf{w}_{is_i}^{\mathrm{T}}\,)^{\mathrm{T}}$ and

$$\mathbf{X}_i = [\,\mathbf{W}_i \mid a_i\mathbf{W}_i \mid b_i\mathbf{W}_i\,].$$

Then we may write

$$Y^{\circ}_{ij} = \mathbf{w}_{ij}\left(\alpha^{\circ} + a_i\alpha^{\mathrm{A}} + b_i\alpha^{\mathrm{S}} + \beta_i\right) + \epsilon_{ij}$$

and

$$\mathbf{Y}^{\circ}_i = \mathbf{X}_i\alpha + \mathbf{W}_i\beta_i + \epsilon_i, \tag{6}$$

where $\epsilon_i = (\,\epsilon_{i1} \quad \ldots \quad \epsilon_{is_i}\,)^{\mathrm{T}}$. Employing the normal-normal conjugacy between likelihood and prior (as in Lindley and Smith, 1972) we have the following complete conditional distributions arising from (6):

for α:

$$\mathcal{N}\left[\left(\sum_i \frac{\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i}{\sigma_i^2} + \mathbf{D}^{-1}\right)^{-1}\left(\sum_i \frac{\mathbf{X}_i^{\mathrm{T}}\mathbf{R}_i^{(\alpha)}}{\sigma_i^2} + \mathbf{D}^{-1}\mathbf{c}\right), \left(\sum_i \frac{\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i}{\sigma_i^2} + \mathbf{D}^{-1}\right)^{-1}\right]$$

where

$$\mathbf{R}_i^{(\alpha)} = \mathbf{Y}^{\circ}_i - \mathbf{W}_i\beta_i\,;$$

and for the β_i:

$$\mathcal{N}\left[\left(\frac{\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i}{\sigma_i^2} + \mathbf{V}^{-1}\right)^{-1}\frac{\mathbf{W}_i^{\mathrm{T}}\mathbf{R}_i^{(\beta)}}{\sigma_i^2}, \left(\frac{\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i}{\sigma_i^2} + \mathbf{V}^{-1}\right)^{-1}\right]$$

where

$$\mathbf{R}_i^{(\beta)} = \mathbf{Y}^{\circ}_i - \mathbf{X}_i\alpha\,.$$

For the σ_i^2 and for $\mathbf{V}$ we again take advantage of conjugacy between likelihood and prior to obtain:

for the σ_i^2:

$$\mathcal{IG}\left[\lambda_1 + \frac{s_i}{2}, \left(\frac{1}{\lambda_2} + \frac{1}{2}\left\|\mathbf{R}_i^{(\sigma^2)}\right\|^2\right)^{-1}\right]$$

where

$$\mathbf{R}_i^{(\sigma^2)} = \mathbf{Y}^{\circ}_i - \mathbf{X}_i\alpha - \mathbf{W}_i\beta_i \text{ and } \|\cdot\|^2 = (\cdot)^{\mathrm{T}}(\cdot)\,;$$

and for $\mathbf{V}^{-1}$:

$$W\left[\left(\sum_i \beta_i \beta_i^{\mathrm{T}} + \rho\Lambda\right)^{-1}, \quad n+\rho\right].$$

Sampling from a Wishart distribution is easily carried out using the method of Odell and Feiveson (1966).

Last, the τ_i and t_i, whose distributions do not enjoy conjugacy with the likelihood, are sampled from discrete twelve- and eight-point distributions respectively. (We label the collapsed upper tail for t_i by the value 8.) In particular, holding all other variables fixed, suppose we consider the exponent of e in the expression $\prod_{j=0}^{s_i} \mathcal{N}\left(Y_{ij}^{\circ} | \mu_{ij}, \sigma_i^2\right)$, from (3), as a function of τ_i alone. Call this function $-g(\tau_i)$. Then the complete conditional distribution for τ_i is

$$\frac{e^{-g(\tau_i)} p(\tau_i)}{\sum_{\tau_k} e^{-g(\tau_k)} p(\tau_k)}, \qquad \tau_k = \frac{1}{4}, \ldots, \frac{23}{4},$$

where $p(\tau)$ is as given in the previous section. Similarly, consider the exponent of e as a function of t_i alone. Call this function $-h(t_i)$. Then the complete conditional distribution for t_i is

$$\frac{e^{-h(t_i)} p(t_i)}{\sum_{t_k} e^{-h(t_k)} p(t_k)}, \qquad t_k = 1, \ldots, 8,$$

where two choices for $p(t)$ were given in the previous section.

3.5. Inverse prediction

One of the determinants of a diagnosis of AIDS or of increased risk of contracting PCP is whether CD4$^+$ cell count has dropped below 200, as cited in section 2. An important question to ask, therefore, is what is the distribution of the time until CD4$^+$ count reaches 200? This is essentially equivalent to asking what is the unconditional latency distribution for AIDS. The method proposed in this section will be employed in the following section to develop such a distribution for self-reported homosexual and bisexual males. Critical values for low CD4$^+$ count other than 200 can of course be used. The resulting estimated distributions emerge as a Bayesian synthesis of previous such estimates (see section 3.3) and the SFMHS data set.

In its simplest form the problem is one of inverse regression (see, for example, Draper and Smith, 1980). Rather than predicting CD4$^+$ counts at a given time we seek to invert this relationship to provide a prediction of time to a specified CD4$^+$ count. The ensuing discussion provides a Bayesian solution to the inverse regression problem for the model in (1)-(4). The distributions we require are for the entire

population of self-reported homosexual males and for the entire population of self-reported bisexual males. Thus we set individual level effects $\beta_i = 0$ and set $b_i = 0$ or 1 accordingly. We seek each estimated distribution for an average subject and thus take a_i equal to the average age in each of the two samples respectively. Other settings for a_i could be used as well. Dropping the subscripts in our model we can thus write the mean CD4$^+$ count $E(Y_0)$ at time Z_0 as

$$E(Y_0) = \mu_0 + \mu_1 Z_0 - \mu_2 (Z_0 - t)^+, \tag{7}$$

where

$$\mu_0 = \alpha_0^o + \alpha_0^A a + \alpha_0^S b\,; \mu_1 = \alpha_1^o + \alpha_1^A a + \alpha_1^S b\,; \text{ and } \mu_2 = \alpha_2^o + \alpha_2^A a + \alpha_2^S b\,,$$

as in (2) but at the population level. Assuming $\mu_1 < 0, \mu_2 > 0$, inversion of relationship (7) produces the desired time for a specified $E(Y_0)$. In particular

$$Z_0 = \begin{cases} (E(Y_0) - \mu_0)/\mu_1 & \text{if } E(Y_0) > \mu_0 + \mu_1 t; \\ (E(Y_0) - \mu_0 - \mu_2 t)/(\mu_1 - \mu_2) & \text{if } E(Y_0) < \mu_0 + \mu_1 t. \end{cases} \tag{8}$$

With a and b specified, for a given $E(Y_0)$ each of the m independent Gibbs replicates may be substituted into (8) producing a collection $\left\{\tilde{Z}_{0k}\right\}_{k=1}^m$. The set of $\tilde{Z}_{0k}$ can be used to construct a kernel density estimate of Z_0. This estimate produces the desired time distribution.

In our case and as in most situations, however, $m = 500$ is insufficient to produce a satisfying kernel density estimate. But, recalling the discussion in section 3.4, to increase m substantially will increase dramatically the amount of additional variate generation. Moreover as observed by Gelfand and Smith (1990) we would do better to utilize the known structure in the model to create a density estimate. That is, we can calculate the complete conditional distribution of Z_0 viewed as a one-to-one function of α_0^o with all other parameters fixed, after which the desired density estimate is obtained as a Monte Carlo integration of this distribution with respect to the m Gibbs replicates. Now a sample size of $m = 500$ will be more than adequate.

More precisely, with $\mu_0' = \mu_0 - \alpha_0^o$ we have

$$Z_0 = \begin{cases} -(\alpha_0^o + \mu_0' - E(Y_0))/\mu_1 & \text{if } \alpha_0^o < E(Y_0) - \mu_0' - \mu_1 t, \text{ i.e. if } Z_0 < t; \\ -(\alpha_0^o + \mu_0' - E(Y_0) + \mu_2 t)/(\mu_1 - \mu_2) & \text{if } \alpha_0^o > E(Y_0) - \mu_0' - \mu_1 t, \text{ i.e. if } Z_0 > t. \end{cases}$$

Recalling that the complete conditional distribution for α_0^o is normal, denoted as $\mathcal{N}\left(\alpha_0 | \mu_{\alpha_0^o}, \sigma^2_{\alpha_0^o}\right)$ for simplicity, the complete conditional distribution for Z_0 will be

$$\begin{aligned} &\mathcal{N}\left(Z_0 | -(\mu_{\alpha_0^o} + \mu_0' - E(Y_0))/\mu_1, \sigma^2_{\alpha_0^o}/\mu_1^2\right) 1_{(-\infty,t]}(Z_0) \\ &+ \mathcal{N}\left(Z_0 | -(\mu_{\alpha_0^o} + \mu_0' - E(Y_0) + \mu_2 t)/(\mu_1 - \mu_2), \sigma^2_{\alpha_0^o}/(\mu_1 - \mu_2)^2\right) 1_{(t,\infty)}(Z_0), \end{aligned} \tag{9}$$

where $1_{(\cdot,\cdot)}(\cdot)$ is an indicator function. The Gibbs replicates determine the parameters of these normal distributions from which the Monte Carlo integration produces a density estimate that is a mixture of the forms in (9).

An alternate approach to solve the Bayesian inverse regression problem is to include a fictitious subject in the sample for whom the only information is the value of a, of b and a particular CD4$^+$ count Y_0. Treating the corresponding Z_0 as an unknown model parameter, we could devise a prior for Z_0. Then implementing the Gibbs sampler as in section 3.4 with the inclusion of this additional subject and with the complete conditional distribution of Z_0 leads to a marginal posterior density estimate for Z_0. The difference between the two approaches is that in one case the specified CD4$^+$ count is treated as a mean (a function of model parameters) while in the other it is treated as a random data point.

4. Results

In this section we present and discuss descriptions of various features of estimated posterior distributions and an investigation of the sensitivity of our conclusions to our prior specifications. We also include a comparison with a previous analysis of a similar data set.

4.1. Posterior distributions for population and individual level parameters

As a starting point we use the lighter-tailed prior $p_1(t)$ for the t_i, with the remaining prior structure as discussed previously. Figure 3 shows the estimated posterior densities obtained for the components of baseline effects α^o, age effects α^A and sexual preference effects α^S. The three rows in Figure 3 are for these three effects, and the three columns are for each of their three components: intercept, pre-change point slope and decrease in slope after the change point. These estimated densities were obtained by mixing the corresponding complete conditional distributions from $m = 500$ independent parallel replications of the Gibbs sampler, each taken to $r = 25$ iterations as mentioned previously. Table 3 is a numerical summary of Figure 3, and gives point estimates in the form of posterior modes and interval estimates in the form of 95% equal-tail credible sets for these population parameters.

The point estimate suggested for the baseline effects is $\widehat{\alpha}^o = (861.0, -87.9, 52.0)$, confirming the basic shape for CD4$^+$ cell decline pictured in Figure 1. Credible sets for all three baseline parameters do not include zero, although the set for α_2^o nearly does, due in part to the much higher variability in its prior. Four of the six credible sets for the remaining α_p^A and α_p^S parameters contain the prior mean zero. Exceptions are an age effect on slope and a sexual preference effect on intercept. The

indicated age-slope effect suggests that the rate of $CD4^+$ cell decline is more rapid for older patients. For instance, the modalvalue for the pre-t_i slope for a 35-year-old self-reported homosexual subject would be $\widehat{\alpha}_1^o + 35\widehat{\alpha}_1^A = -114.6$. The indicated sexual preference-intercept effect suggests that those who identified themselves as bisexual upon study entry had a higher overall $CD4^+$ cell count at seroconversion than self-reported homosexuals at seroconversion. An estimate of this increase is given by $\widehat{\alpha}_0^S = 113.0$. The lower estimated intercept for the homosexual group may reflect effects of comorbidities not explicitly measured in the study. This result suggests a potential difference in the progression from seroconversion to AIDS between self-reported homosexuals and bisexuals; we discuss this point further at the end of this section.

Pearson correlations between population parameters may be obtained from the replicates generated by the Gibbs sampler. Doing this we obtain Table 4. The correlations in our prior structure had exactly the signs exhibited in this table and magnitudes of 0.5. The posterior correlation between α_0^S and α_1^S is the only one that has moved very far away from its prior value. We did compute estimated bivariate posterior distributions for some of the α pairs by mixing the appropriate bivariate normal complete conditional densities (not shown), but found that these displays added little to our understanding.

We summarize individual level parameters in Figure 4 by showing the three possible bivariate scatterplots of intercepts and slopes among the β_{0i}, β_{1i}, and β_{2i}. The plotted points are estimated posterior means computed as simple averages of the appropriate Gibbs samples for each of the 331 individuals in the study. The range of intersubject variability is large for the intercepts (roughly 2200 units, or 2.5 times $\widehat{\alpha}_0$), very large for the pre-t_i slopes (roughly 750 units, or 8.5 times $\widehat{\alpha}_1$), and enormous for the post-t_i changes in slope (roughly 500,000 units, or 10,000 times $\widehat{\alpha}_2$). It is somewhat surprising that our procedure was able to detect any population level signal given the vagueness of our prior and the presence of such noise. The extreme variability in the β_{2i} seems due to the little available information to estimate features of the post-change point portion of the curve with at most five observations available per subject. The change point seems to occur very late in the measurement period (see findings for the t_i and τ_i below), making post-change point slope estimation difficult or nearly impossible on an individual basis without additional data at subsequent time points. Compounding the problem are censoring mechanisms at later time points for subjects who, for reasons unrecorded, discontinued their visits. The scatterplots show a rather high degree of association between the β_{0i} and the β_{1i}, as subjects with higher than average intercepts are associated with more steeply decreasing curves. No significant correlation is evident between the β_{2i} and either the β_{0i} or β_{1i}, no

doubt due mainly to the general difficulty in measuring the post-t_i individual slopes.

Figure 5 provides histograms of the estimated posterior means for the remaining individual level parameters τ_i, t_i, and $\sigma_i, i = 1, \ldots, n$. As with the β_i, these estimates are simple averages of the corresponding $m = 500$ iterates. It is this averaging that produces rather continuous estimated posteriors for the τ_i and t_i even though both had discrete priors. The $\widehat{\tau}_i$ histogram indicates a much more even distribution of times since seroconversion than was present in the rather peaked prior distribution taken from Bachetti and Moss (1989); the data have revised our prior to suggest a more uniform distribution of seroconversion times for the subjects in our study. The picture for $\widehat{t}_i$, on the other hand, is less diffuse than the prior and is skewed to the left. For most subjects, the CD4$^+$ cell level seems to begin to decrease more steeply between 5 and 5.5 years after seroconversion. However, there is also a significant minority of subjects for whom the switch to the second regime occurs much sooner, perhaps as early as 3 years after seroconversion. Finally, the histogram of $\widehat{\sigma}_i$ values shows that most subjects have estimated values much less than the prior mean of 200, with only a very few σ_i estimated to exceed 300. However, the presence of a few very large $\widehat{\sigma}_i$ values suggests that our decision to assume heterogeneous variances was worthwhile.

The modal growth curves for self-reported homosexual and for self-reported bisexual subjects are shown in Figure 6, with change points at the respective estimated posterior medians. We see that the population level curve for the former group lies below that of the latter and has a larger decrease in slope. We also show boxplots of the $\widehat{t}_i$ for each group. Comparing these boxplots, the self-reported homosexual group has a more dispersed distribution, exhibiting considerable skewness to the left with median roughly three months later than that of the more homogeneous self-reported bisexual group. Thus it turns out that most of the subjects with relatively early estimated change points are found in the self-reported homosexual group.

Graphical displays of the results of applying our Bayesian solution for the inverse prediction problem outlined in section 3.5 are provided in Figure 7. These two curves are the estimated posterior distribution of the mean number of years from seroconversion to a CD4$^+$ count of 200 for homosexuals and bisexuals of average age, 36.3 and 35.1 years respectively at study entry. In mixing the densities of the form given in (9) we took $t = E(t_i) = 5.1$, the prior mean under $p_1(t)$. Our analysis suggests a slightly higher mean time from seroconversion to AIDS for self-reported bisexual subjects (a point estimate of 6.3 years vs. 6.1 years) and consistent with our finding that progression patterns for self-reported bisexuals had higher estimated CD4$^+$ counts at seroconversion. There is also a bit more variability

in the picture for this group, suggested for instance by comparing the credible sets. This is probably due to a smaller sample size (67 bisexual vs. 264 homosexual). Similar inverse prediction analysis could of course be conducted if one desired comparisons of subjects at ages other than group-specific average ages at study entry.

4.2. Prior sensitivity and a brief comparison with a previous analysis

To examine the robustness of our results we considered perturbations of key elements in our prior specification whose elicitation was less precise. This is an important issue in any Bayesian analysis, and especially so in this problem where intersubject variability abounds. (It could have turned out that the data contained essentially no information about the overall population parameters, and that our prior distributions were driving results completely.) We did not perturb the prior on τ_i, nor the prior on α^o because of fairly accurate external information as cited previously. We also did not perturb the deliberately vague prior specification for the σ_i^2.

We were concerned, however, about the potential impact of increased levels of uncertainty in the prior variance-covariance matrices for the population mean effects α^A and α^S as well as for the β_i. Unanticipated additional uncertainty could have led to all components of α^A, α^S or both being statistically indistinguishable from zero. To decrease precision we replaced $\mathbf{D}^A$ by $4\mathbf{D}^A$, $\mathbf{D}^S$ by $4\mathbf{D}^S$ and Λ by 4Λ. This amounted to doubling the uncertainty in each of these prior components. As expected, these modifications did affect the related posterior distributions, but no dramatic changes were observed. In particular, the components of α^A and α^S that emerged as significantly different from zero in the original analysis, α_1^A and α_0^S, continued to have this property: alternate 95% credible sets became $(-1.42, -0.25)$ and $(47.37, 232.13)$ respectively. In fact, both of the related point estimates actually moved *further away* from zero ($\widehat{\alpha}_1^A = -0.79$ and $\widehat{\alpha}_0^S = 134.39$). For the components of the β_i the alternate scatterplots (not shown) had an appearance very similar to those in Figure 4, over roughly the same range for the β_{0i} and β_{1i} and about twice the range (10^6 units) for the β_{2i}. Histograms for the $\widehat{\tau}_i, \widehat{t}_i$ and $\widehat{\sigma}_i^2$ did not change appreciably, nor did the inverse prediction posterior for the self-reported homosexual group. The corresponding distribution for self-reported bisexuals did widen somewhat from that given in Figure 7. The alternate 95% credible set was roughly 0.5 years wider, but the point estimate was essentially unchanged at 6.27. Overall, we conclude that these changes in prior specifications do not bring about substantial changes in our results.

Next, since the prior for t_i was built piecemeal from several sources we considered the changes brought about by switching to the heavier-tailed prior $p_2(t)$ for the t_i. Not surprisingly, this led to a histogram of $\widehat{t_i}$ values having a higher mode (up to roughly 5.9 years) but the same left-skewed appearance as in Figure 5b, the range of values now going from 3.5 to 6.5. A more interesting result is that the alternate credible set for α_0^A of $(-0.07, 5.00)$ came very close to excluding zero. The corresponding interval in Figure 3d does as well, suggesting an age-related contribution to the overall model intercept.

We also analyzed the model with both of the above perturbations to the prior in place. The changes observed were essentially the logical union of those described in the preceeding two paragraphs. In particular, the population parameters had somewhat more diffuse estimated posteriors, but those that were significantly different from zero previously remained so. The inverse prediction credible sets were a bit wider, being up to 1.2 units wider for self-reported homosexuals and about 2.1 units wider for self-reported bisexuals. Figure 7 lost some of its symmetry for this latter group, showing instead a heavier upper tail. Finally, the ranges for β_{0i} and β_{1i} were roughly the same as in Figure 4, but the β_{2i} range became even larger (up to 2.5×10^6). Overall, these analyses show that our results are fairly robust to imprecision in prior specifications.

Finally, we give a brief comparsion of our work with that of De Gruttola et al. (1990). As mentioned in section 2, these authors fit a considerably simplified version of our model to a different portion of the SFMHS data set, excluding all incomplete cases. They fit a random-effects model with a likelihood similar to that given in (3), but assuming only a simple straight-line growth curve without a random change point t_i. They also did not allow for the age and sexual preference covariates, nor for nonhomogeneous subject variances σ_i^2. Their decision to conduct only a complete case analysis is perhaps the most restrictive, as it eliminated almost 40% (130 subjects) of the 331 HIV-positive subjects available. To calibrate their results for HIV-positive subjects, De Gruttola et al. also included HIV-negative patients with complete data records in their analysis, setting $Z_{ij}^* = 0, j = 1, \ldots, s_i$, for these subjects. These authors adopted a parametric empirical Bayes approach to their analysis and hence did not place priors on α, V or on their common σ^2 but rather estimated these from the marginal distribution of the data. They did however assume that the τ_i represent errors in the Z_{ij}, as have we, and took their distribution to be a seven-point discrete distribution derived from a coarser version of the infection-time distribution calculated by Bachetti and Moss (1989). Before fitting their model, De Gruttola et al. (1990) transformed the Y_{ij}° counts to the square root scale in order to stabilize the observational variance

across individuals. This we did not do, because our assumption of nonhomogeneous variances makes it unnecessary and because all of our prior information related to the original scale. We have no idea what form of growth curve would be appropriate on the square root scale.

With all of the above differences in model and data only very rough reconciliation between the analyses is possible. Using the square root scale De Gruttola et al. obtained an estimated population intercept of 33.15 and an estimated population slope of -2.13. Since the intercept corresponds to the square root of population CD4$^+$ count at seroconversion, its square, 1099, may be compared with estimated intercepts from our study. In particular, these are 939 for a 35 year old (average aged) self-reported homosexual and 1052 for a self-reported bisexual of the same age. Since $dY/dZ = d\sqrt{Y}/dZ \div d\sqrt{Y}/dY$, a slope of -2.13 corresponds to a rate of change at seroconversion of -141 on the original scale. From Figures 3, 4 and 5, using model values we find a pre-t_i slope of -113 and a post-t_i slope of -156 for a 35 year old self-reported homosexual. Similarly, we find a pre-t_i slope of -130 and a post-t_i slope of -146 for a 35 year old self-reported bisexual.

5. Conclusions and Discussion

We have conducted a fully Bayesian analysis of the progression of HIV infection using longitudinal CD4$^+$ counts by employing a high-dimensional hierarchical model. Our approach accommodates individual piecewise-linear growth curves with random unobserved change points, heterogeneous variances, unbalanced and incomplete data, several population level covariates and unobserved infection times. These extensions of classical Bayesian or frequentist parametric growth-curve models have been possible through recent advances in Bayesian computations for high-dimensional models, in particular stochastic relaxation using the Gibbs sampler. To our knowledge, this sampling-based approach is the only feasible way to analyze a model involving all of the preceding features. This computationally-demanding technology has yielded further understanding of the post-infection behavior of CD4$^+$ counts for the San Francisco Men's Health Study cohort, as summarized in the preceding section. We note in addition that though certain of our 95% credible sets include zero, directional adjustments, perhaps using modal values, could be made when reporting point estimates.

Our inferences provide useful additional information for clinicians, AIDS hospice workers, counsellors and health policy experts. As more data become available from this cohort beyond the fifth occasion of observation we can both validate our model and sharpen our predictions if necessary. Further data on CD4$^+$ counts may possess richer structure than our present data set: the individual time series may

be longer and additional population covariates, possibly time-varying, may be available in other studies. There may also be multivariate outcomes at each occasion, e.g. longitudinal p24 antigen levels and $CD4^+$ counts together. Under such circumstances several directions for generalization of our modeling approach suggest themselves, including more general patterns on variance-covariance matrices such as auto-regressive process models for larger numbers of longer time series, the luxury of fewer parametric assumptions on the shapes of the individual level curves, and generalizations of the growth-curve models used here to accommodate multivariate outcomes.

References

Bacchetti, P. and Moss, A. R. (1989). Incubation period of AIDS in San Francisco. *Nature*, 338, 251–253.

Bowen, D. L., Lane, H. C. and Fauci, A. S. (1989). Immunopathogenesis of the acquired immunodeficiency syndrome. *Annals of Internal Medicine*, 103, 704–709.

Brookmeyer, R. and Gail, M. (1987). Biases in prevalent cohorts. *Biometrics* 43, 739–749.

Brookmeyer, R. and Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* 45, 325–335.

Centers for Disease Control (1989). Guidelines for prophylaxis against *Pneumocystis carinii* pneumonia for persons infected with human immunodeficiency virus. *MMWR* 38, S-5.

Collier, A. C., Bozzette, S., Coombs, R. W. et al. (1990). A pilot study of low-dose zidovudine in human immunodeficiency virus infection. *New England Journal of Medicine* 323, 1015–1021 (October 11).

Cox, D. R. and Solomon, P. J. (1986). Analysis of variability with large numbers of small samples. *Biometrika* 73, 543–554.

De Gruttola, V., Lange, N., and Dafni, R. (1990). Modeling the progression of HIV infection. Submitted for publication.

Devash, Y., Calvelli, T. A., Wood, D. G. et al. (1990). Vertical transmission of human immunodeficiency virus is correlated with the absence of high-affinity / avidity maternal antibodies to the gp120 principal neutralizing domain. *Proc Natl Acad Sci*, 87, 3445–3449.

Draper, N. R. and Smith, H. (1980). *Applied Regression Analysis*, 2nd edition. New York: Wiley.

Eyster, M. E., Goedert, J. J., Sarngadharan, M. G. et al. (1985). Development and early natural history of HIV-III antibodies in persons with hemophilia. *Journal of the American Medical Association*, 253, 2219–2223.

Eyster, M. E., Ballard, J. O., Gail, M. H., Drummond, J. E. and Goedert, J. J. (1989). Predictive markers for the acquired immunodeficiency syndrome (AIDS) in hemophiliacs: Persistence of p24 antigen and low T4 cell count. *Annals of Internal Medicine*, 110, 963–969.

Eyster, M. E., Gail, M., and Ballard, J. (1987). Natural history of Human Immunodeficiency Virus infection in hemophiliacs: Effects of T-cell subsets, platelet counts, and age. *Annals of Internal Medicine*, 107, 1–6.

Fahey, J. L., Taylor, J. M. G., Detels, R. et al. (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. *New England Journal of Medicine*, 322, 166–172 (January 18).

Fauci, A. S., Macher, A. M., Longo, D. L. et al. (1984). Acquired immunodeficiency syndrome: epidemiologic, clinical, immunologic, and therapeutic considerations. *Annals of Internal Medicine*, 100, 92–106.

Fearn, T. (1975). A Bayesian approach to growth curves. *Biometrika*, 62, 89–100.

Fischl, M. A., Parker, C. B., Pettinelli, C. et al. (1990). A randomized controlled trial of a reduced daily dose of zidovudine in patients with the acquired immunodeficiency syndrome. *New England Journal of Medicine* 323, 1009–1014 (October 11).

Geisser, S. (1970). Bayesian analysis of growth curves. *Sankhyā, Series A*, 32, 53–64.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. To appear, *Journal of the American Statistical Association.*

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Kalbfleisch, J. D. and Lawless, J. (1988). Estimating the incubation period for AIDS patients. *Nature (London)*, 333, 505–.

Lagakos, S. W. and De Gruttola, V. (1989). The conditional latency distribution of AIDS for persons infected by blood transfusion. *Journal of Acquired Immunodeficiency Syndromes*, 2, 84–87.

Lang, W., Perkins, H., Anderson, R. et al. (1989). Patterns of T lymphocyte changes with human immunodeficiency virus infection: from seroconversion to the development of AIDS. *Journal of Acquired Immunodeficiency Syndromes*, 2, 63–69.

Lange, N. and Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, 84, 241-247.

Lee, J. C. and Geisser, S. (1972). Growth curve prediction. *Sankhyā, Series A*, 34, 393–412.

Leibovitz, E., Rigaud, M., Pollack, H. et al. (1990). *Pneumocystis carinii* pneumonia in infants infected with the human immunodeficiency virus with more than 450 CD4 T lymphocytes per cubic millimeter. *New England Journal of Medicine*, 323, 531–533 (August 23).

Leoung, G. S., Feigal, D. W., Montgomery, A. B. et al. (1990). Aerosolized pentamidine for prophylaxis against *Pneumocystis carinii* pneumonia. *New England Journal of Medicine*, 323, 769–775 (September 20).

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Lifson, A., Hessol, N., Rutherford, G., et al. (1989). The natural history of HIV infection in a cohort of homosexual and bisexual men: Clinical manifestations, 1978–89. Abstract, *Fifth International Conference on AIDS*, Montreal, Canada.

Longini, I., Clark, W., Byers, R. H. et al. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, 8, 831–843.

Lui, K. J., Lawrence, D. N., Morgan, W. M. et al. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proc Natl Acad Sci*, 83, 2913–2917.

Masur, H., Ognibene, F. P., Yarochoan, R. Y. et al. (1989). CD4 counts as predictors of opportunistic pneumonias in human immunodeficiency virus infection. *Annals of Internal Medicine*, 111, 223–231,

Medley, G. F., Anderson, R. M., Cox, D. R. and Billard, L. (1987). Incubation period for of AIDS in patients infected via blood transfusion. *Nature (London)*, 328, 719–721.

Odell, P. L. and Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, 61, 198–203.

Phair, J., Muñoz, A., Detels, R. et al. (1990). The risk of *pneumocystis carinii* infection among men infected with human immunodeficiency virus type 1. *New England Journal of Medicine*, 322, 161–165 (January 18).

Rao, C. Radhakrishna (1987). Prediction of future observations in growth curve models (with comments). *Statistical Science*, 4, 434–471.

Taylor, J., Fahey, J., Detels, R. and Giorgi, J. (1989). CD4 percentage, CD4 number, and CD4:CD8 ratio in HIV infection: Which to choose and how to use. *Journal of Acquired Immunodeficiency Syndromes*, 2, 114–124.

Volberding, P. A., Lagakos, S. W. et al. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine*, 322, 941–949 (April 5).

Table 1. Prior distribution for the unobserved random infection times. Values for $p(\tau)$ in the table are $\times 10^4$.

τ	1/4	3/4	5/4	7/4	9/4	11/4	13/4	15/4	17/4	19/4	21/4	23/4
$p(\tau)$	685	1089	1371	1415	1356	1238	1191	959	414	172	55	6

Table 2. Prior distributions for the unobserved random change points.

t	1	2	3	4	5	6	7	≥ 8
$p_1(t)$	0.05	0.1	0.1	0.15	0.15	0.15	0.1	0.2
$p_2(t)$	0.0375	0.075	0.1125	0.1125	0.1125	0.1125	0.075	0.4

Table 3. Posterior modes and 95% credible sets for the population parameters.

(a) posterior modes

covariate effects	intercept	pre-changepoint slope	decrease in slope at the changepoint
baseline	861.0	-87.9	52.0
age (in years)	2.2	-0.76	-0.23
sexual preference*	113.0	-17.1	-27.4

(b) 95% credible sets

covariate effects	intercept	pre-changepoint slope	decrease in slope at the changepoint
baseline	(776.2, 950.9)	(-102.5, -74.2)	(6.1, 89.2)
age (in years)	(-0.23, 4.8)	(-1.3, -0.28)	(-1.5, 0.76)
sexual preference*	(34.8, 196.0)	(-41.8, 5.5)	(-71.7, 14.2)

* self-reported homosexual (coded 0) or bisexual (coded 1) at study entry.

Table 4. Pearson correlations between population parameters.

	baseline	
	α_1	α_2
α_0	-0.524	-0.377
α_1	-	0.503

	age	
	α_1^A	α_2^A
α_0^A	-0.542	-0.367
α_1^A	-	0.550

	sexual preference	
	α_1^S	α_2^S
α_0^S	-0.809	-0.394
α_1^S	-	0.557

Figure 1. Piecewise-linear expected individual $CD4^+$ cell counts over time, with unknown change point t_i.

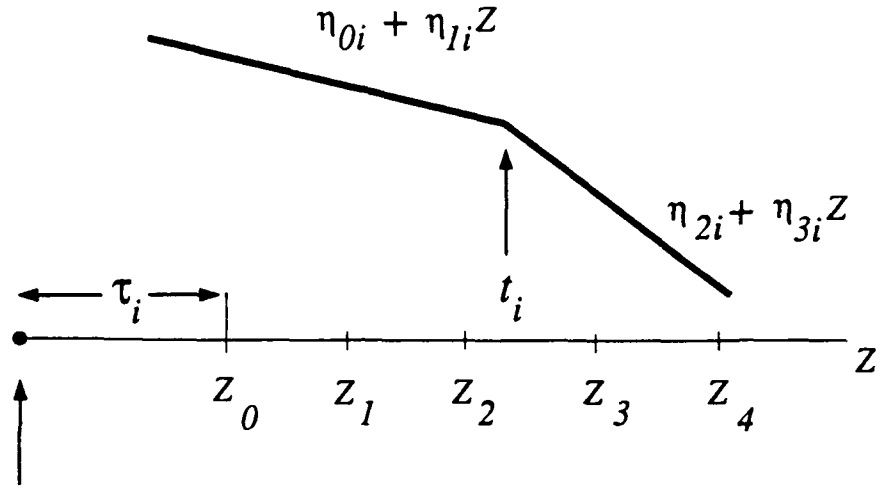


Figure 2. A directed graph for the hierarchical Bayes model.

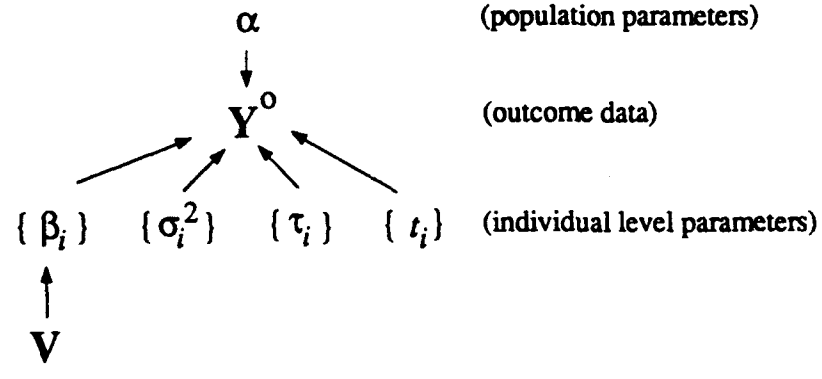


Figure 3. Estimated posteriors, population parameters

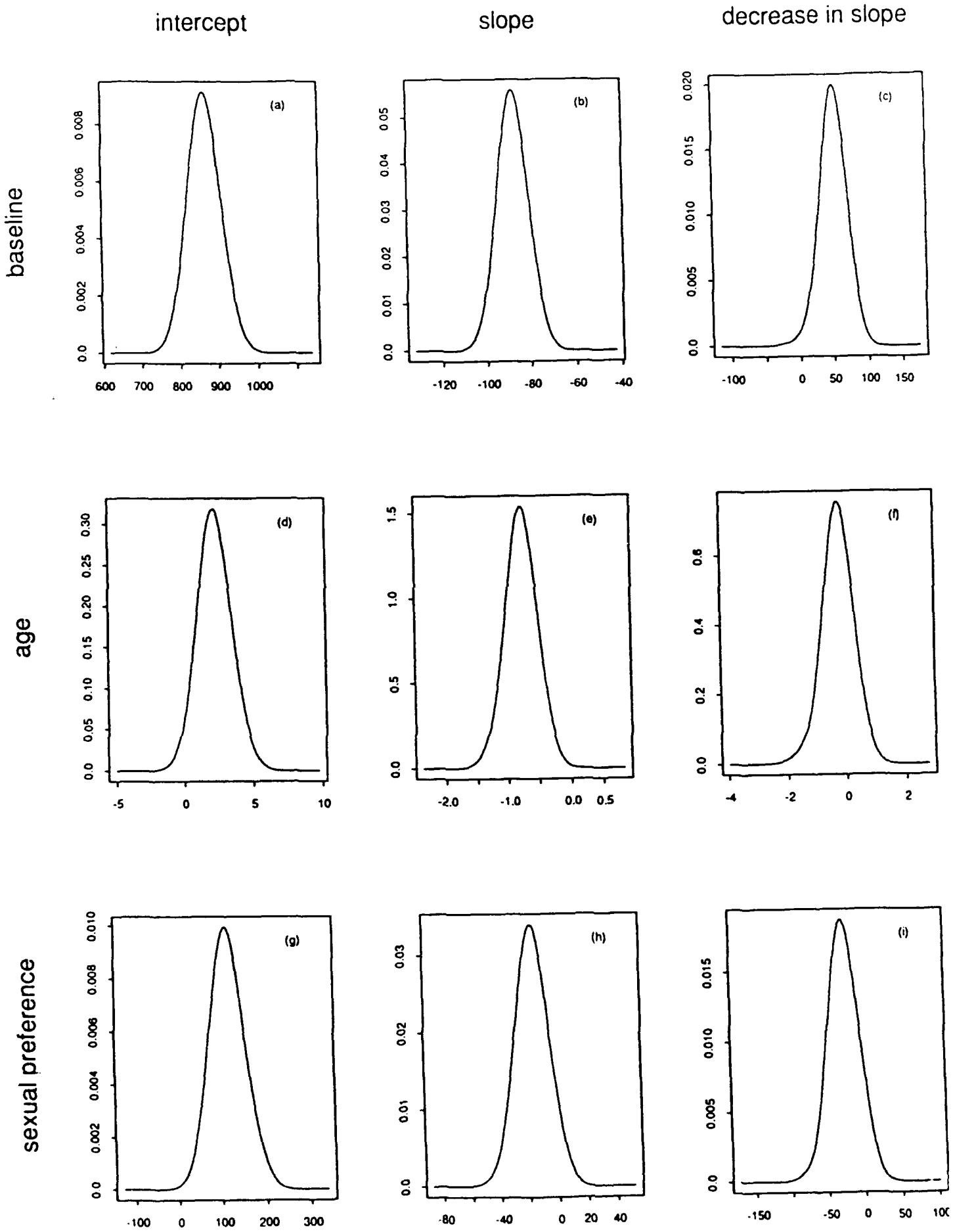


Figure 4. Scatterplot matrix, components of beta vector

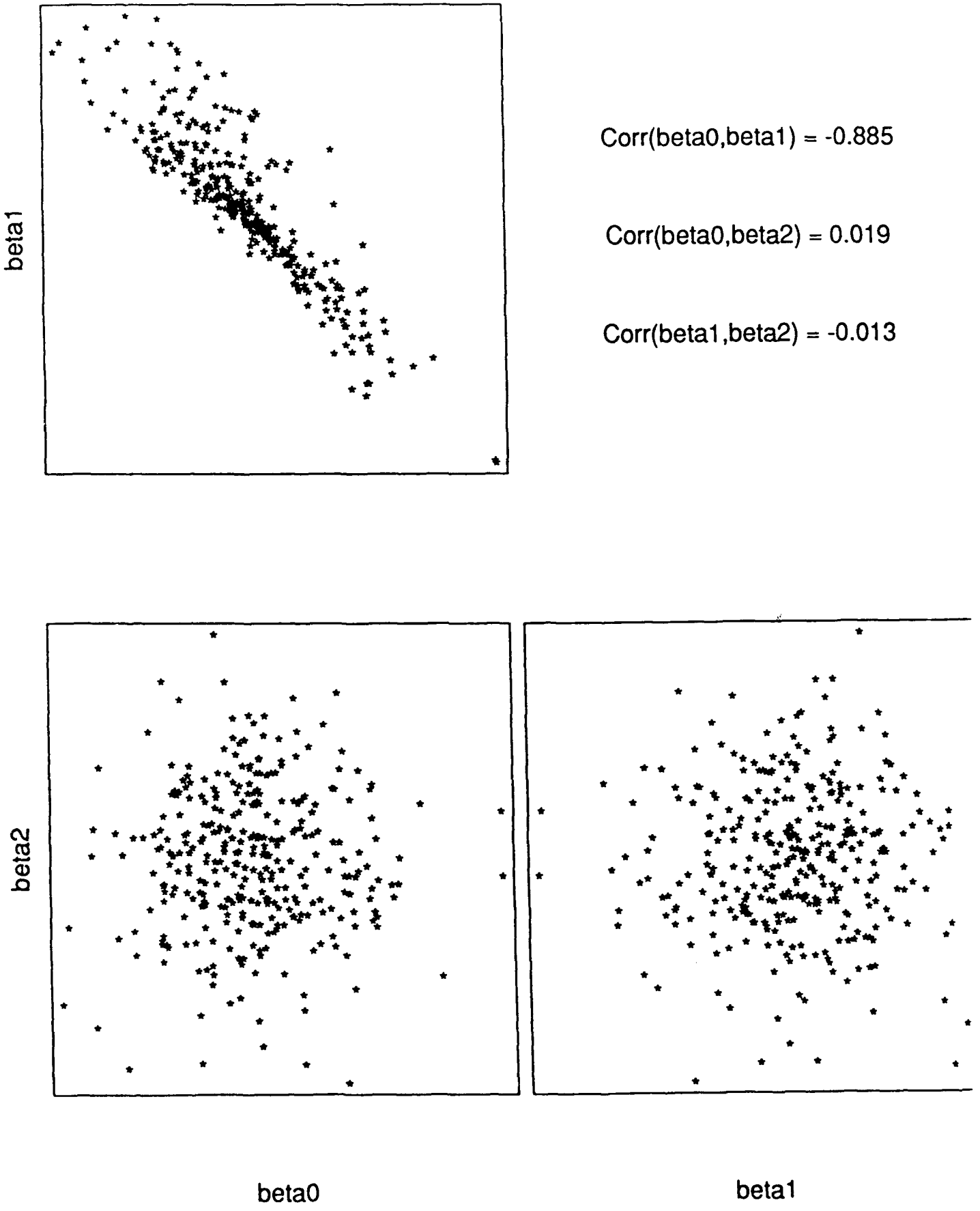
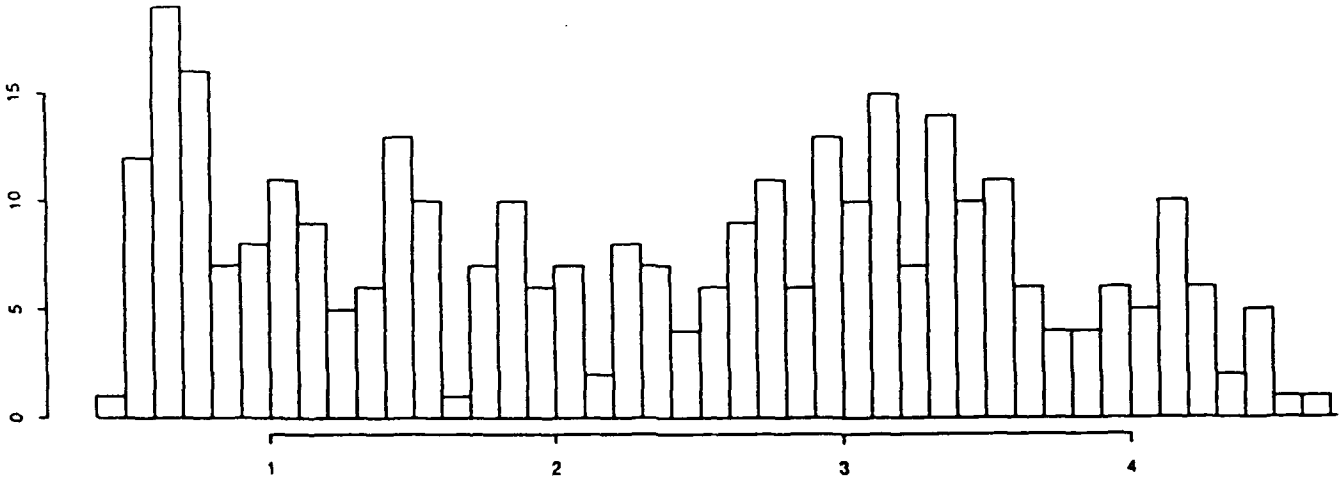
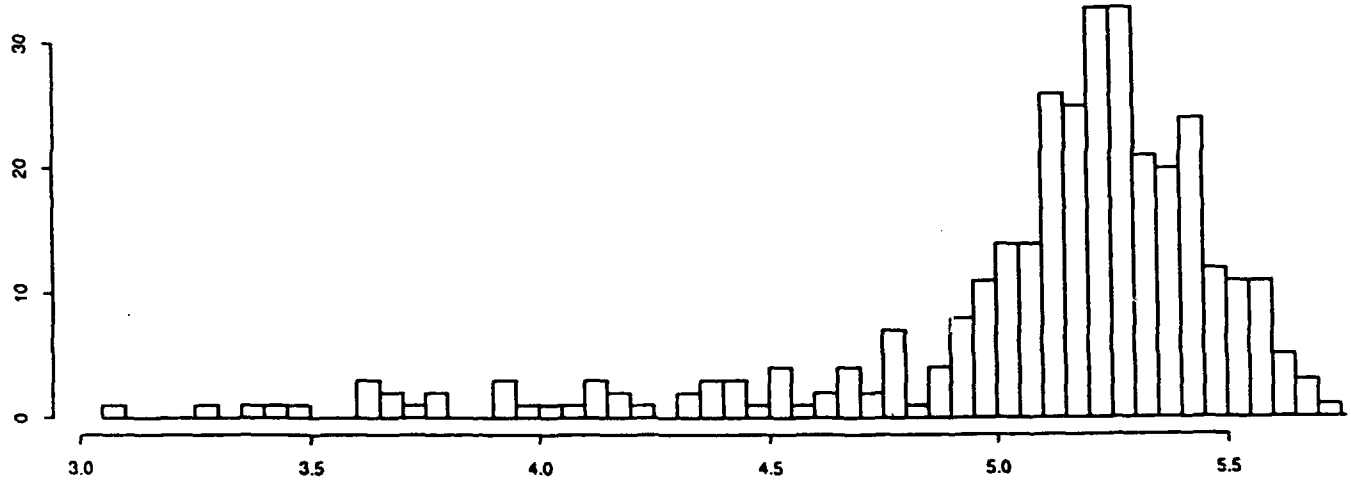


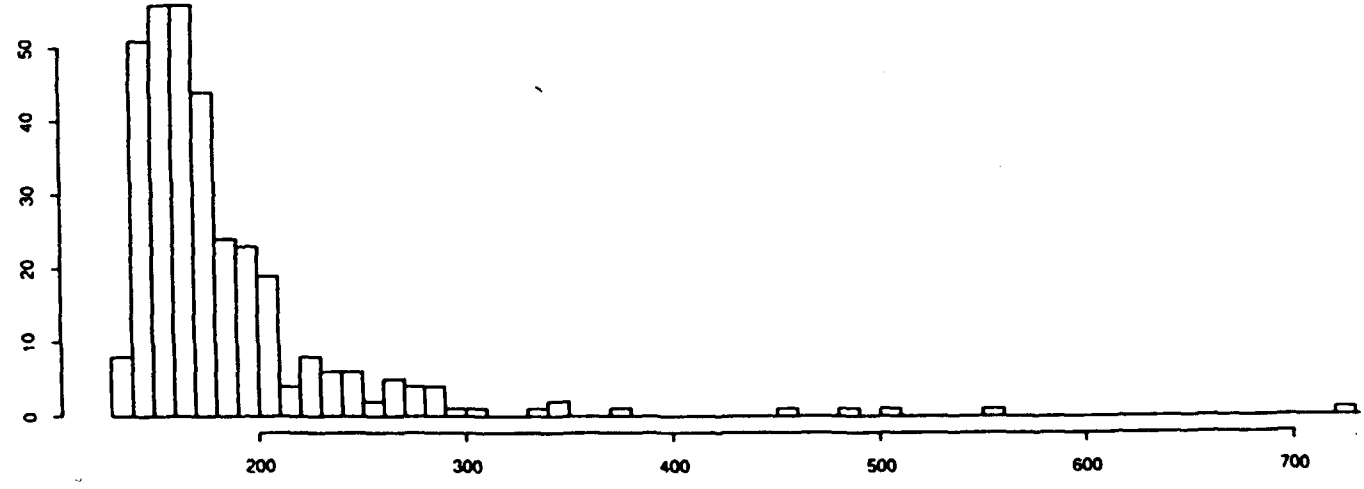
Figure 5. Histograms of other individual level parameters



a) Histogram of tau i estimates



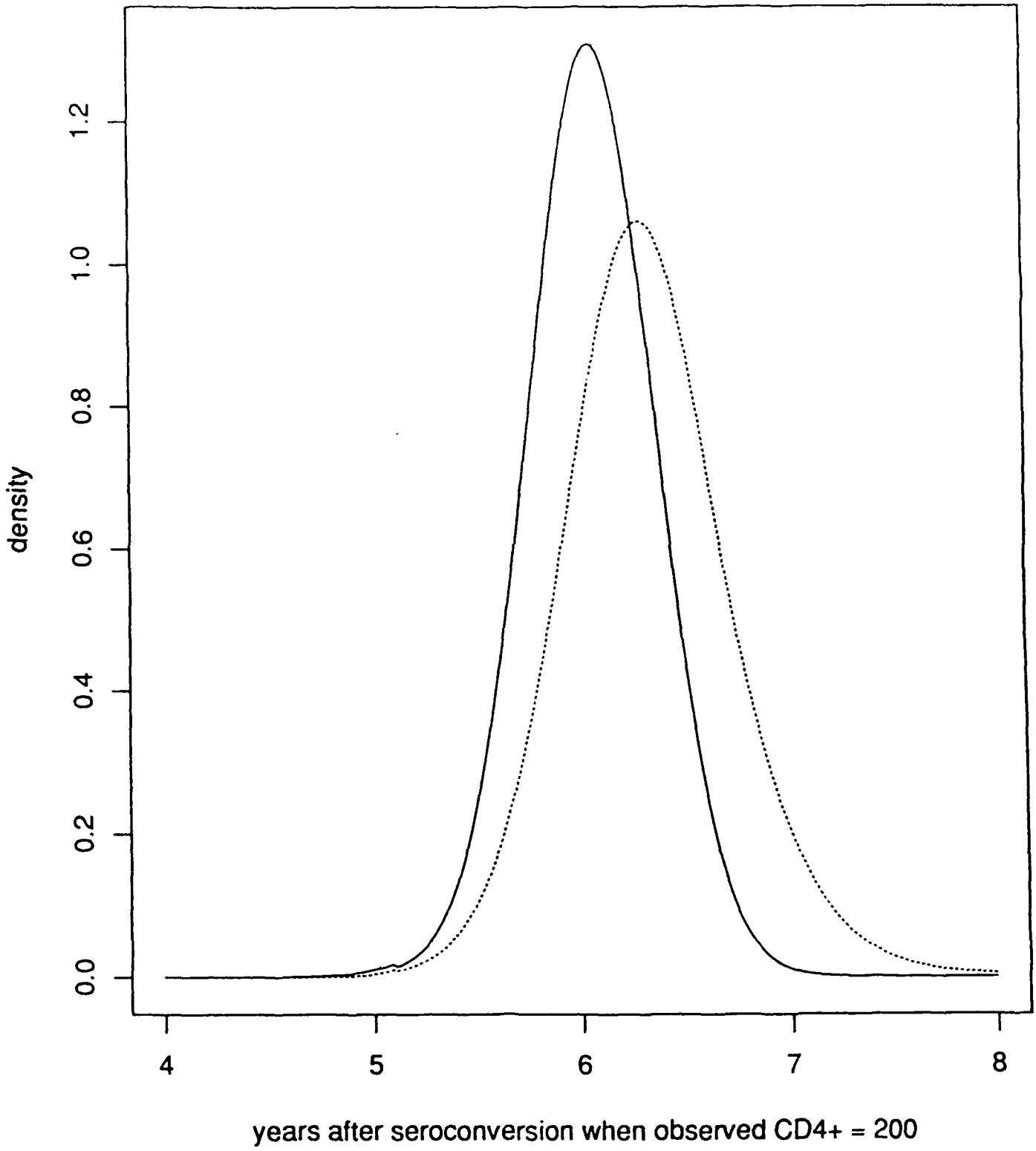
b) Histogram of t i estimates



c) Histogram of sigma i estimates

Figure 6. Fitted population model with estimated changepoint distributions

Figure 7. Estimated posteriors, inverse prediction



Solid line: Homosexual group; mode is 6.07 ; 95% CS is (5.44 , 6.66)
Dashed line: Bisexual group; mode is 6.31 ; 95% CS is (5.56 , 7.19)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 461	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Hierarchical Bayes Models for the Progression of HIV Infection using Longitudinal $CD4^+$ Counts		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Nicholas Lange, Bradley P. Carlin, Alan E. Gelfand		8. CONTRACT OR GRANT NUMBER(s) N0025-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 111		12. REPORT DATE November 27, 1992 13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

KEY WORDS: AIDS, Gibbs sampler, growth curves, heterogeneity, inverse prediction, marginal posterior distribution, prior specification, random change points, sexual preference.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Taking the absolute number of $CD4^+$ cells (also known as T helper cells, T4 cells, and CD4 cells) as a marker of disease progression for persons infected with the human immunodeficiency virus (HIV) we model longitudinal series of such counts for a sample of 331 subjects in the San Francisco Men's Health Study. We conduct a careful and fully Bayesian analysis of these data. We are able to employ individual level nonlinear models incorporating critical features such as incomplete and unbalanced data, population covariates, unobserved random change points, heterogeneous variances, and errors-in-variables. Using results of previously published work from several different sources we construct rather precise prior distributions. Our analysis provides marginal posterior distributions for all population parameters in our model for this cohort. Using an inverse prediction approach we also develop the posterior distributions of time for $CD4^+$ count to reach a specified level.